**METODY EKSPLORACJI DANYCH**

Laboratorium. Modele logitowe. Regresja logistyczna.

Prowadzący: Wykonali:

dr inż. Romuald Hoffmann pchor. Michał ADAMCZEWSKI

pchor. Mikołaj ADAMSKI

pchor. Przemysław SUJECKI

Zadanie 1.

W wybranej losowo grupie studentów jednolitych studiów magisterskich z warszawskich uczelni badano ich stan cywilny w zależności od roku studiów. Wyniki obserwacji zebrano w tabeli 1. Przyjęto przez M oznaczać osoby będące w związku małżeńskim, natomiast przez W – osoby stanu wolnego.

Obraz zawierający tekst, numer, linia, Czcionka

Opis wygenerowany automatycznie

W zadaniu proszę:

1. Wyznaczyć zależność stanu cywilnego badanych studentów do roku studiów, zakładając najpierw liniowy model prawdopodobieństwa, a następnie model logitowy.

2. Na podstawie opracowanych modeli i przeprowadzonych obliczeń sformułować własne wnioski.

3. Wyniki analizy proszę zawrzeć w postaci sprawozdania, do którego proszę dodać jako załączniki wszystkie pliki z obliczeniami (obliczenia można przeprowadzić w dowolnie wybranym narzędziu.

Funkcja logistyczna jest nieliniową funkcją jednej zmiennej, którą z reguły jest zmienna 𝑡 odpowiadająca czasowi. Funkcja logistyczna, której wykresem jest krzywa przypominająca kształtem literę „S”, należy do rodziny krzywych sigmoidalnych (ang. sigmoid curve).

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

Przyjmijmy do dalszych rozważań, że zapiszemy jako:

Obraz zawierający tekst, Czcionka, linia, biały

Opis wygenerowany automatycznie

Jak łatwo zauważyć, funkcja ta nie jest liniowa względem parametrów. Ideą metody Hotellinga jest takie przekształcenie funkcji logistycznej, aby uzyskać zależność liniową, w której występowałyby pewne proste funkcje parametrów.

Różniczkując funkcję logistyczną względem zmiennej objaśniającej otrzymujemy:

Obraz zawierający Czcionka, linia, pismo odręczne, tekst

Opis wygenerowany automatycznie

Stąd otrzymujemy:



Pamiętając o tym, że przyrost zmiennej objaśniającej t (np. w funkcji trendu) może być przyjmowany jako równy jedności mamy 𝑑𝑦 = Δ𝑦 (to wynika z założenia, że różniczka funkcji jest równa jej przyrostowi). Wówczas mamy:



Oznaczamy przez: przez , stąd

Wówczas ostatnie zależność ta przyjmuje postać:



Parametry 𝛽0 i 𝛽1 w powyższej zależności możemy oszacować metodą najmniejszych kwadratów, ponieważ jest to zależność liniowa względem parametrów. W ten sposób oszacujemy parametr 𝛽0 = 𝑎 logistycznego modelu oraz parametr 𝛽1 = , jako funkcję parametru 𝑘. Oznaczmy przez 𝛽̂0 i 𝛽̂1 oceny

parametrów 𝛽 i 𝛽1. Wówczas mamy:



W celu oszacowania parametru 𝑏 przekształcimy wzór:

Obraz zawierający Czcionka, linia, numer, tekst

Opis wygenerowany automatycznie

w następujący sposób:



Z powyższego mamy:



Całkiem naturalne wydaje się wiec przyjęcie za ocenę parametru ln(𝑏) takiej wartości, która dla oszacowanych wartości parametrów 𝑎 i 𝑘 (tzn. 𝑎̂ oraz 𝑘̂) minimalizuje wyrażenie:

Obraz zawierający Czcionka, linia, biały, diagram

Opis wygenerowany automatycznie

Otrzymujemy wówczas następujący estymator ln(𝑏):

Obraz zawierający Czcionka, tekst, linia, biały

Opis wygenerowany automatycznie

Jeżeli chcielibyśmy otrzymane metodą Hotellinga wzory na estymatory parametrów 𝑎, 𝑏 oraz 𝑘 zapisać za pomocą danych, to przyjmą one postać jak poniżej.

Biorąc pod uwagę, że mamy oraz zbiór obserwacji {𝑦𝑡∶ 𝑡 = 1, … , 𝑛} to mamy:

Obraz zawierający Czcionka, tekst, linia, biały

Opis wygenerowany automatycznie

Parametr a szacujemy ze wzoru:

Obraz zawierający Czcionka, linia, tekst, diagram

Opis wygenerowany automatycznie

Oszacowanie parametru 𝑘 ma postać:

Obraz zawierający Czcionka, tekst, linia, pismo odręczne

Opis wygenerowany automatycznie

Parametr b szacujemy ze wzoru:

Obraz zawierający Czcionka, tekst, linia, biały

Opis wygenerowany automatycznie

Modele logitowe stosujemy głownie w modelowaniu zmiennych jakościowych, inaczej zwanych dychotomicznych, które opisują fakt wystąpienia lub niewystąpienia analizowanego zjawiska; np. zakupu określonej rzeczy, korzystania/rezygnacji z określonej usługi, itd.

Wartość oczekiwana zmiennej objaśnianej może być interpretowana jako warunkowe prawdopodobieństwo realizacji danego zdarzenia przy ustalonych wartościach zmiennej objaśniającej lub zmiennych objaśniających (może być ich więcej). Wartość 𝑦̂𝑡 uważamy za oszacowanie prawdopodobieństwa wystąpienia intersującego zdarzenia. Wadą tego rozwiązania jest to, że praktycznie wartość 𝑦̂𝑡 może być spoza przedziału wartości [0, 1] i wobec tego użycie takiego modelu daje wątpliwy efekt. Aby temu zaradzić dokonuje się odpowiedniego przekształcenia prawdopodobieństwa z przedziału

[0, 1] w przedział (−∞, +∞) stosując tak zwaną transformację logitową postaci:

gdzie:

𝑝 – jest prawdopodobieństwem wystąpienia zdarzenia.

Iloraz nazywany jest szansą wystąpienia zdarzenia.

Prawdopodobieństwo 𝑝 szacujemy na podstawie zaobserwowanych danych dotyczących zmiennych 𝑌 i 𝑋. W praktyce empirycznej wartość prawdopodobieństwa 𝑝 szacujemy jako częstości empiryczne 𝑝𝑡 wystąpienia zdarzenia. Wobec tego dla każdego wystąpienia mamy transformację:

Stosując przekształcenie odwrotne do transformacji 𝐿 = mamy:

Prawdopodobieństwo 𝑝 szacujemy na podstawie zaobserwowanych danych dotyczących zmiennych 𝑌 i 𝑋. W praktyce empirycznej wartość prawdopodobieństwa 𝑝 szacujemy jako częstości empiryczne 𝑝𝑡 wystąpienia zdarzenia. Wobec tego dla każdego wystąpienia mamy transformację:

1. Wykres zawieranych małżeństw.

Obraz zawierający tekst, zrzut ekranu, Wielobarwność, diagram

Opis wygenerowany automatycznie

Wykres przedstawiający liczbę osób w zależności od roku studiów, będących zaręczonymi lub wolnymi.

## Model liniowy

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

MODEL LOGITOWY

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie





|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rok studiów | Liczba studentów | Wolni | Małżeństwo | Pi | 1-Pi | ln(p(t))/(1-p(t)) |
| 1 | 6 | 5 | 1 | 0,166666667 | 0,833333 | -1,609437912 |
| 2 | 6 | 5 | 1 | 0,166666667 | 0,833333 | -1,609437912 |
| 3 | 5 | 2 | 3 | 0,6 | 0,4 | 0,405465108 |
| 4 | 5 | 1 | 4 | 0,8 | 0,2 | 1,386294361 |
| 5 | 6 | 1 | 5 | 0,833333333 | 0,166667 | 1,609437912 |
| Suma | 28 | 14 | 14 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rok studiów | (at+b) | e^-(at+b) | 1+e^-(at+b) | 1/1+e^-(at+b) | p(t) |
| 1 | -1,9062 | 6,727475757 | 7,72747576 | 0,129408365 | 0,129408 |
| 2 | -0,9292 | 2,532482381 | 3,53248238 | 0,283087045 | 0,283087 |
| 3 | 0,0478 | 0,953324433 | 1,95332443 | 0,511947725 | 0,511948 |
| 4 | 1,0248 | 0,358868232 | 1,35886823 | 0,735906526 | 0,735907 |
| 5 | 2,0018 | 0,135091899 | 1,1350919 | 0,880985937 | 0,880986 |

Współczynniki:

Obraz zawierający tekst, Czcionka, zrzut ekranu, biały

Opis wygenerowany automatycznie

|  |  |
| --- | --- |
| a = | 0,19503 |
| b = | -0,07813 |

Współczynniki:

Obraz zawierający tekst, Czcionka, biały

Opis wygenerowany automatycznie

|  |  |
| --- | --- |
| a = | 0,977 |
| b = | -2,88325 |

Budowanie modeli do predykcji:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Predykowane wartości dla modelu liniowego i logitowego:

Obraz zawierający tekst, numer, menu

Opis wygenerowany automatycznie

**Wnioski:**

Z pierwszego wykresu możemy wywnioskować, że im późniejszy rok studiów, tym więcej osób zawiera małżeństwa, zgodnie z dodatnim współczynnikiem a.

Ogólnie rzecz biorąc, możemy zauważyć, że model logitowy trochę lepiej radzi sobie z predykowaniem wartości. Np. dla obserwacji nr 19 widzimy, że predykcja dla modelu liniowego wynosi 0.70199, a dla modelu logitowego 0,73590.

Porównując współczynnik R2, dla modelu liniowego wynosi 0,3204, a dla modelu logitowego wynosi 0,3879.

Nie posiadamy zbyt wielkiej ilości obserwacji i zmiennych, przez co modele nie będą zbyt dobrze predykowały wartości.

Bibliografia:

[0] dr inż. Romuald Hoffmann, prof. WAT, Notatki dla studentów, Warszawa 2023.

[1] Larose D. T., Metody i modele eksploracji danych, Wydawnictwo Naukowe PWN, Warszawa 2012.

[2] <https://scikit-learn.org/stable/>

[3] Fred Nwanganga, Mike Chapple, Praktyczne uczenie maszynowe w języku R, Wiley, Warszawa, 2022.

[4] Laurence Moroney, Sztuczna inteligencja i uczenie maszynowe dla programistów, Helion O’Reilly, 2021.

[5] <https://pl.wikipedia.org/wiki/Wikipedia:Strona_g%C5%82%C3%B3wna>

[6] Aurelie Geron, Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow, Helion O’Reilly, 2018